UCRL- MI-145050

ASCI White GPFS status: August 2001

Richard Hedges, William Loewe, Tyce Mclarty, David Fox, Mark Grondona, Robin Goldstone, Terry Heidelberg

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

August 14, 2001

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

This report has been reproduced
directly from the best available copy.

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information
P.O. Box 62, Oak Ridge, TN 37831
Prices available from (423) 576-8401
http://apollo.osti.gov/bridge/

Available to the public from the
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd.,
Springfield, VA 22161
http://www.ntis.gov/

OR

Lawrence Livermore National Laboratory
Technical Information Department’s Digital Library
http://www.llnl.gov/tid/Library.html

ASCI White GPFS status: August 2001

This report by the Scalable I/O Project (SIOP) covers the progress of parallel I/O performance for the GPFS parallel file system on the ASCI White computer system.

The major hardware and software configuration changes which affect GPFS performance are: (1) The Colony switch adapters have been upgraded from single/single to double/single; and (2) the system software has been upgraded from Mohonk (PSSP 3.2) with GPFS 1.3 used during February 2001 testing to the present Mohonk2 (PSSP 3.3) with GPFS 1.4.

The tests discussed here were performed on white using up to 300 compute nodes and the GPFS file systems using 16 dedicated I/O nodes (servers).

The tests performed utilize the POSIX interface to GPFS. The noted system changes to white have improved both GPFS peak read and write by about a factor of two for the large scale (>= 128 nodes) tests.

Machine characteristics

As of February 2001, white nodes were connected to the Colony switch by single/single adapters. By late July 2001, those adapters had been upgraded to double/single switch adapters and the supporting software (PSSP 3.3 and GPFS 1.4) was installed.

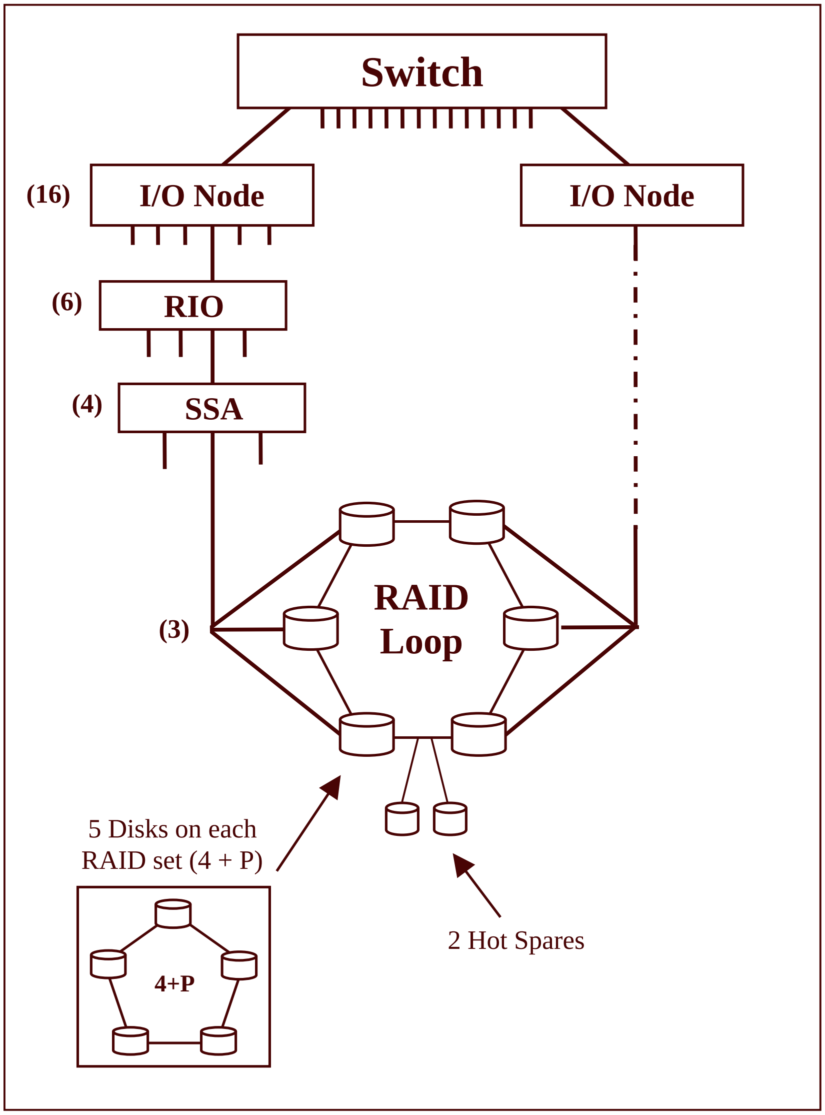
White has 512 total nodes including 16 dedicated GPFS I/O server nodes. The nodes are 16-way SMP nodes comprised of 375Mhz Power3 processors. Each of the 16 SSA servers has 6 RIO drawers with 4 SSA adapters per drawer. Each of these 24 adapters is connected to an SSA loop, which consists of 32 disks spread across two SSA enclosures where each loop contains 6 4+P arrays and 2 hot spares. [See Figure 1.] An individual server is responsible for serving 3 of the six disks in a loop, thus there are a total of 1152 disks available:

(3 disks/loop * 24 loops/node * 16 nodes = 1152 disks)

Due to file system limitations, however, the /p/gw1 file system actually consisted of only 1024 disks during the February testing. To achieve this reduction in disks, 1 disk per server was removed from the /p/gw1 file system for every third SSA loop. This resulted in the removal of 8 disks per server, or 128 disks total, from the file system for a final size of 1024 disks.

During the first set of tests on February 17, the /p/gw1 file system being tested consisted of only 1024 disks. As of July 28, it consists of all 1152 disks. The impact of this configuration difference is assumed to be minimal: There is more than adequate disk bandwidth beyond the throughput capacity of the switch.

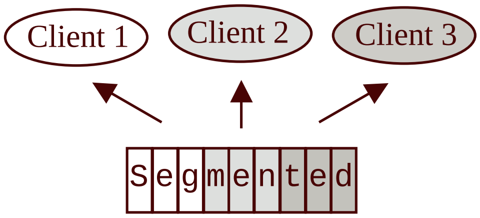
Figure 1 diagrams the hardware layout for the /p/gw1 file system. The numbers in parentheses to the left in the figure show the number of components at each level.



[Fig. 1] *File System Hardware*

Test Description

These tests use the POSIX version of the IOR test code with segmented data pattern and transfer size = GPFS block size. This is the most basic and familiar test used by the SIOP to probe peak I/O performance of the GPF S file systems on the ASCI machines. The pattern is segmented when each process performs I/O from/to a large contiguous area of the file distinct from other processes. [See Figure 2.]



[Fig. 2] *Segmented Pattern*

More explicitly we use: segmented pattern; 1 client process per node; varying nodes; create, write and read a single common file (size = 512 *MB***n*). Each process accesses 512*MB*, which are contiguous in the file where for each process *p* in 0...n -1, *p*'s first byte is located at location *p**512*MB*. The size of each individual transfer is the GPFS block size = 512*KB*.

Results

February 17, 2001: single/single, IP

Nodes	**30**	**60**	**128**	**256**
MB/s write	**2473**	**3061**	**2170**	**2653**
MB/s read	**2589**	**3285**	**2306**	**3007**

July 28, 2001: double/single, KLAPI

Nodes	**32**	**64**	**128**	**300**
MB/s write	**4262**	**5202**	**4793**	**5094**
MB/s read	**4862**	**4779**	**6921**	**6967**

[Note: Different node counts for February and July tests.]

University of California
Lawrence Livermore National Laboratory
Technical Information Department
Livermore, CA 94551